DISCUSSION PAPER FOR BREAKOUT SESSION III

STATE OF THE ART IN ISSUES OF UNCERTAINTY AND VARIABILITY FOR PBPK MODEL APPLICATIONS

Frédéric Y. Bois
INERIS
Parc Alata BP 2
F-60550 Verneuil en Halatte
France
Tel: +33 3 44 55 65 96
Fax: +33 3 44 55 61 75
Email: frederic.bois@ineris.fr

Abstract

Issues of uncertainty and variability are of significant importance to risk assessment. This paper describes briefly the past and current practices devised to deal with those issues when making predictions with PBPK models, the strengths and limitations of those practices; the currently considered "best" practices – and under what conditions they should be used – and finally what is needed (*e.g.*, research, data, software, reporting/documentation standards, *etc*.) to further improve and expand the use of such models in the future.

1. Introduction

Models are devices (mathematical, physical, biological…) that are easy to manipulate and "act like" the real world in some aspects we are interested in. On those aspects they also tend to give answers faster than simple observation. A primary use of models is to make predictions. Being able to forecast death by toxicity, for example, we are better prepared to avoid it. In a more elaborated use, model predictions can also be used to test theories. A prediction invalidated by subsequent observation is the telltale that the theory or analogy on which the model is based is flawed. In an even more sophisticated use, models can give us access to information on parameter values (when hidden and hard to measure). This also, typically requires some predictions of easy to measure quantities for which data exists and that are sensitive to the parameter of interest. Care must be taken in that case if other parameters (of the model or not) also influence the prediction made or the data observed. The final degree of sophistication closes the loop between models and experiments: Models can be used for experimental design, *i.e.*, to make predictions on the data best suited to estimate some parameters.

PBPK models are mechanistic and have an extended domain of applicability (usually much larger than that of classical pharmacokinetic models). Given that property, PBPK modeling has so far been targeted mainly to purely predictive use, but we are witnessing an evolution toward the other uses, with the development of statistical tools able to deal with complex models. That is true for both toxicity risk assessment and therapeutic drug development. A short survey of the various applications of PBPK models points to the following:

A major purpose, estimation of internal doses: It is well known that the same external dose can lead to very different internal doses and effects depending on the species, timing, or individual (see therapeutic drugs with narrow therapeutic window) (Andersen 1995). Further, internal dose is not necessarily proportional to external dose, and that complicates dose-response relationships. PBPK models are able to predict internal doses and can therefore explain (or point to the lack of explanation) for the above phenomena, which can all have an impact on risk assessments.

Extrapolations: Those are also purely predictive applications of PBPK modeling. They were used early in toxicology and risk assessment, with the work of the Dow Chemicals toxicology group (e.g., Reitz *et al.* 1980), teasing out non-linearity in dose-response relationships. On the other hand, Bishoff and Dedrick (Bischoff *et al.* 1971) were more interested in inter-species extrapolation for anti-cancer agents dosing. The various extrapolations found in the literature concern:

- *Exposure routes and schedules*: Oral intake, for example, is not equivalent to pulmonary intake. After oral intake the absorbed substance goes through the gut wall and liver where extensive metabolism usually takes place. That is the origin of the well-known first-pass effect. While *via* the lung, the substance gets the opportunity to disperse in all organs before getting to sites to metabolism. Various routes of exposure have been PBPK-modeled (oral (Bischoff *et al.* 1971), inhalation (Frederick *et al.* 1998), intra-venous, dermal (Poet *et al.* 2000), intra-muscular, intra-cerebral, intra-peritoneal…), most of time with usually valid simplifying assumptions (*e.g.*, peritoneal injection being assimilated to an intra-liver injection, *etc*.). Those should however be critically assessed when reviewed.
- *Dose levels*: Non-linearity between external and internal dose is required to justify any effort in PBPK modeling for dose extrapolation, but such modeling can be an added value because of the other possibilities of extrapolation it offers (Reitz *et al.* 1980).
- *Time*: PBPK models, even if linear, are quite cumbersome to solve analytically. They are practically always treated numerically and their input (external dose) can be easily made an arbitrary function of time. They can therefore be used to extrapolate between short-term and long-term exposures, or between single and repeated exposures. Recently, time-varying parameter models (describing a

growing person, or pregnant woman, for example) have been proposed to improve the quality of such extrapolation over long time periods (Gentry *et al.* 2003) (See the section below on inter-occasion variability for the use of time evolving-parameters to model inter-occasion differences).

- *Space*… What are the pharmacokinetics of acetaminophen in microgravitational conditions? PBPK modeling can help explore this question (Srinivasan *et al.* 1994). The question is now more about drug treatment for space missions, but it could apply to toxics, in a century or two… (Note: that's a reminder to all those who use "body weight" in their models instead of "body mass").

- *Species*: For the same external dose, internal dose may vary in nature or quantity between animal species. PBPK models break down that extrapolation problem by splitting it into the transposition of many parameters of various types (anatomical, physiological, substance-specific, metabolic, *etc*.).

- *Individuals*: PBPK models can be used to assess inter-individual differences in internal dose and explain their origins. Anatomical, physiological, and metabolic (genotypic or phenotypic) differences can be described. The coupling with multilevel population models allows inference to be made about those differences on the basis of individual data. Obviously, differences in internal dose condition at least partly differences in susceptibility to toxicity and an understanding of their origins can help design more effective regulations. Individuals can be grouped by sub-populations as in (Mezzetti *et al.* 2003).

- *Occasions*: It seems that little attention has been paid so far to inter-occasion (or intra-individual) variability in environmental risk assessment. The picture is somewhat different for therapeutic drugs. PBPK models can account for a time-dependence of their parameters (*e.g.*, periods of rest and work-load, pregnancy, lifetime history) (Jonsson and Johanson 2001), but the field and thinking about the potential applications is still in infancy.

High throughput screening: That is the most important use for the pharmaceutical industry, besides inter-species extrapolation. Pharmacokinetics are an obvious experimental bottleneck to *in silico* drug development (Leahy 2006). Substance-independent PBPK models, in which the substance-specific parameters are computed using predictive chemistry, QSAR/QSPR or *in vitro* methods, offer a fast computerized way to screen libraries of more or less virtual chemicals. Extensive-coverage regulations, such as REACH in Europe, will also require high-throughput hazard and risk assessment. PKPK models could also have a role to play in those contexts and regulatory uses could develop quickly.

Parameter identification: Seems that this is one of the games in physical science, but that we are not yet there. Maybe one day… Actually, using a PBPK to better derive a carcinogenic potency value is an exercise of the sort. But it's not a PBPK model parameter that is sought. Seeking an inter-individual variance is more akin to specific.

Hypothesis Testing and Theory Checking: Now that we have tools to calibrate PBPK models and assess the minimal distance between a model's predictions and some data, we should be ready to focus on testing and improving their structural assumptions. Hypothesis testing, in the form "is model A preferable to model B", requires a solid statistical apparatus, because it is a frightening perspective to reject a reasonable model in favor or a bad one on the basis or flimsy experimental evidence. An (imperfect) example of such use for of PBPK modeling can be found in (Bois *et al.* 1991). The optimal choice of statistical method (*e.g.*, log-likelihood, predictive likelihood, information criteria, *etc.*) is still an open question on which guidance would be useful, but in any case, it is likely that there is no universal and definitive best choice.

Dose reconstruction: Interestingly, I have seen through the years several attempts (in the form of submitted articles) to reconstruct dose from internal biomarker data *via* PBPK modeling, but how many ended being published? (Georgopoulos *et al.* 1994). Dose reconstruction is a pure inverse problem and requires a somewhat strict statistical treatment, and could have important regulatory or legal applications.

Experimental Design: Actually, designing experiments is much easier than analyzing them (Bois *et al.* 1999). That is not yet common regulatory use, but could well expand, because of ethical and regulatory requirements to spare animal lives, minimize human volunteer exposures and risks *etc*.

Among the above applications, only a few are deeply concerned with uncertainty and variability and are of significant importance to risk assessment as performed nowadays. Section 2 will focus on uncertainty issues and section 3 on variability. Ancillary matters of quality assurance (section 4) and reporting (section 5) are treated next. We end with a general discussion of those issues.

2. Uncertainty issues in regulatory applications

Issues of uncertainty and variability are almost always entangled, because early and simple statistical models of data processes usually lump together parametric and model uncertainty with variabilities of all sorts (inter- and intra-individual variability in particular). I choose here to separate them, because they should not be confused, and partly for ease of exposition. Partly only, because if uncertainty arises naturally in primary (“1^{st} moment”) predictions or inferences, it also contaminates inference we conduct about variability (“2^{nd} moment” applications in which there is often a large degree of uncertainty about the value of the estimated variances).

Past and current practices

We will not go into the details of the various ways (use of parametric or Monte Carlo simulations, classical likelihood-based methods and more recently Bayesian approaches) that have been or are currently in use to assess parametric uncertainty. Those have been covered by the previous chapters.

Past uses of parametric uncertainty analyses in regulatory applications have been about as scant as the use of PBPK models for such purposes. An interesting point is actually that (to my knowledge) the first and only direct use of a PBPK model in a passed regulation (methylene chloride ruling, Occupational Safety and Health Administration (OSHA) 1997) was triggered by the possibility to perform an associated (Bayesian) uncertainty analysis. Indeed, a PBPK model with a handful of nonlinear differential equations and some twenty parameters must appear like a tar pit of uncertainty to the lay public (to which regulation must appear legitimate and clear, to the possible extent). And while we may be able to argue that body mass is about 73 kg (but then the variability issue creeps in), it is much more difficult to make a convincing argument that V_{max} for benzene is exactly 3.14159 in mice. Amazingly, by the way, it does not seem difficult to make believe that it scales to the power 3/4 (an exact ratio) across species, but that is the exception to a general rule: People trust exact numbers only in accounting. In risk assessment, betting is the name of the game, and I think that anyone dealing with uncertainty, probabilities, risks, wants to have an idea of the range of possibilities. Actually, most people do not have a clear idea of what a probability or risk is. Most people understand that it's some fuzzy number and want the size of the ballpark before coming in (an issue we will touch upon again below in a section devoted to communication of results). So, OSHA used parametric uncertainty analysis to estimate the distribution of internal dose estimates, which were used in input to a standard cancer model. That actually amounted to relaxing the unrealistic (even if convenient) hypothesis of a binomial distribution of error when fitting a standard multistage model. Why should all individuals in a similarly exposed group have the same probability of cancer, when we know that internal dose vary and that it is not known precisely for each?

Model uncertainty is another problem. PBPK models have little uncertainty about their anatomical or physiological structure, but large uncertainties can still subsist about the necessary degree of tissue lumping, metabolism, or proper definition of internal dose, for example. Here again, classical and Bayesian methods for model comparison, model checking (Akaike criterion, cross-validation…), and combining/weighting/averaging predictions from different models, have been covered in previous chapters [hopefully]. I have seen very little discussion of uncertainty in PBPK model structure in a regulatory context. I suspect again that regulators do not like to be on shifty grounds, and model uncertainty seems (and is) an even more fundamental problem than parametric uncertainty. So, to a large extent model uncertainty is unacceptable or taboo to many. It may also be an issue that goes beyond the realm of statistics, into the fuzzier area of heuristics, as far as mechanistic models are concerned. The closest to a discussion of model structure I have seen in a regulatory application was the comparison (albeit informal) of Clewell et al. (2000) *vs*. Fisher et al. (Abbas and Fisher 1997; Bois 2000; Fisher 2000) model for trichloroethylene. The point has been resolved for now by going to a third, unified, model (Hack *et al.* 2006). Actually the statistician's usual defense to a model uncertainty attack is to shift to nonparametric methods. There are, for example, nonparametric versions of hierarchical population models (Davidian and Gallant 1992; Rosner and Mueller 1997). With those, we do not need to posit a standard (*e.g.*, lognormal) distribution of parameter values among a population of subjects. However, I have not yet seen those applied to PBPK modeling. Yet, no one seems to have yet seriously or at least publicly discussed the need for nonparametric approaches to PBPK modeling.

Current best practices

Given the very embryonic stage of the art in the statistical treatment of PBPK modeling, it seems difficult to prescribe definite *best practices*. Yet, it is clear that an evaluation of parametric uncertainty affecting model predictions is in order. The Bayesian framework is currently *en vogue* for clear, quite easy to grasp, and coherent parametric uncertainty analyses of PBPK models. It can coherently integrate experimental data and more fuzzy expertise to sharpen the distribution of poorly known parameters and it outputs statistical distributions for model predictions of any kind. Were the uncertainty about all model parameters (or about a subset of "sensitive parameter", as defined by an adequate sensitivity analysis) well known, simple Monte Carlo simulations should suffice. They also output distributions of predictions. Issues of presentation of complex outputs such as (multivariate) distributions are discussed below.

A note on "adequate" sensitivity analysis is in order. Two cases may arise. First, when only prior information on parameters is available (or considered), a simple Monte Carlo analysis may point to unacceptably large uncertainties in model predictions. We may then want to find which are the most "sensitive" parameters, *i.e.*, those that condition the most the output distributions. For such a task a global sensitivity analysis (Spear *et al.* 1991; Saltelli *et al.* 2000) of the model predictions considered is recommended. The second case is when we want to select a subset of parameters to calibrate, for example *via* a Bayesian analysis. This requires determining the parameters to which the predictions *of the data at hand* are sensitive. Then using a conditional independence argument it can be justified to set the insensitive parameters to fixed values (*e.g.*, a central location estimate) because the data likelihood would also be insensitive to them, and hence the posterior of the sensitive parameters. *Ergo*, inference from the data would be unchanged. Obviously if the parameters pointed to as sensitive for data predictions are not sensitive for the application sought (*e.g.,* a risk estimate) one may question the usefulness of using the data for the problem at hand. This can be turned inside out into an experimental design question: "What new data would be useful for sharpening my predictions (under the assumption that my model is more or less correct)?" (Müller 1999).

It is recommended, before doing "out of the hat" predictions, to check the model by comparing its outputs with corresponding data *not used* for calibration or other form of parameter setting (Zhang 1993; Gelman *et al.* 1996). In that "cross-validation" exercise, the criteria used to judge model adequacy should be defined *a priori* (*i.e.*, conceived as a hypothesis test). If those criteria are fulfilled and the model judged adequate, it is possible to recalibrate the model with all available data included, and that should further reduce uncertainty about parameter estimates.

As to model uncertainty, the best to be afforded for now involves a clear definition of the model's domain of applicability (a term which should be preferred to the more commonly used "domain of validity"), a discussion of modeling hypotheses made and their possible alternatives (even if deemed less likely).

Needs

Many possibilities are open for improvement of PBPK modeling in the realm of uncertainty analysis:

- I would first advocate a more extensive use of decision analytic concepts in risk assessment. For example, now that we have reasonable tools for uncertainty analysis, we could start asking questions such as: What is the level of uncertainty below which a firm regulatory decision is acceptable? What level would trigger interim (precautionary) decisions and further research and data collection? *Etc.*
- Uncertainty about model structure is largely neglected, mostly by lack of research, tools, and standards. We could develop guidelines on domain of applicability issues: How to define, check, specify, and describe the domain of applicability of a PBPK model? There is also a need for sensitivity analyses of modeling assumptions (not limited to the statistical model part, as in Bayesian statistics, but extended to the metabolic description, for example). Some interesting thoughts on "model understanding" have been proposed by A. Gelman (Gelman 2004) and could be further explored. It could be worth exploring semiparametric or nonparametric approaches for both statistical population models and mechanistic PBPK models (in particular in "open", many-possibilities areas such as metabolic pathways). Good and pragmatic guidance on the use model comparison methods should be developed, with a focus on PBPK modeling.

- The same way that reference values are proposed by the ICRP, we might want to go one step further and propose reference statistical distributions for given populations. Those could be used as default values in standard Monte Carlo simulations or as priors in more sophisticated statistical analyses. We might want to have them provided as good analytical approximations of the real (probably mixture) distribution, for ease of use in simulations (yet, simulation tools should adapt to the real world complexity, not the converse).
- Software is scarce for those analyses. Either we code everything ourselves in Mathematica®, Matlab®, Splus®, R or other generalist packages, or we use somewhat cryptic and research-oriented specialized software such as BUGS® or *MCSim*®, or we wait for our favorite special-purpose commercial package to implement the tools we need.

3. Variability issues

In a simple world, where we just want to limit the risks borne by the "reference man", it makes sense to think of "the" carcinogenic potency q for compound X. The true value for q is not know with precision and to protect ourselves against that uncertainty we may chose to use its upper 95% confidence limit (q^*) in regulatory applications. When we realize that we are dealing with real people, we can still continue to use q^* as if nothing happened and consider that we are protecting *almost* everybody. But the problem is that the unprotected fraction of the population may well be identifiable (*e.g.*, with the help of a PBPK model) and may well have a say about regulations. Shouldn't all members of the population, including the "sensitive" ones, be offered (if not guaranteed) the same level of protection? At the same time, focusing risk abatement efforts on those, hopefully few, individuals, could also lead to more efficient regulations… But how far should go the search for sensitive subjects? When we are using animal data, should we also consider variability between individual animals? between species? By the way, we seem to be ready to spend a lot of research time and money to understand interspecies differences in terms of mechanisms, and to leave inter-individual differences in the gray box of random differences. Why? Medicine is going toward the individualization of therapies, do we want to regulate at the same level? The technical aspects of understanding, defining, modeling and estimating variability have been covered in the previous chapters, and we wish to call the attention of the reader to above questions, which are essential to rational decision-making and regulation.

Whatever their extent, variability issues are with us, and it is essential to be able to disentangle then from uncertainty. As mentioned earlier, early statistical models aggregated uncertainty and variability. It was enough to report average measurements taken at the same time from various subjects, and conduct pharmacokinetic analyses on such data. Under some models that was "sufficient" statistics, and that saved valuable space in paper publications. Decisive work by Sheiner et al. (1992) showed that that is flawed and the same went on to propose multilevel ("population") statistical models able to estimate separately variability and uncertainty (including uncertainty in estimates of variability…). The original applications used classical pharmacokinetic models, but it turned out that PBPK models were also amenable to the same treatment (albeit in a Bayesian numerical framework where complexity was less of an issue) (Gelman *et al.* 1996). Earlier work had also explored those issues in a PBPK context, but with firm *a priori*, non-inferential, distributions (Bogen and Spear 1987; Droz *et al.* 1989; Droz *et al.* 1989).

The gist of population models is to identify and isolate the elements (parameters) affected by variability, either *a priori* (by defining distributions of values that reflects population variability and carrying out Monte Carlo simulations) (Gearhart *et al.* 1993), or by regression of model-predictable pharmacokinetic data (indirect evidence) to the parameter level. The later requires some form of statistical modeling and inference (e.g., hierarchical population modeling) (Sheiner and Ludden 1992). But sophisticated modeling is not *per se* a worthy goal. The most important is to identify population sub-groups that are at risk and to be predictive of variability expected in a population to be effectively protected. An important goal is to identify easily measurable covariates that explain (or indirectly correlate with) variability and potentially to susceptibility. It is all well and good to find that CYP2E1 correlate with susceptibility to compound X, but do we need phenotyping of the whole US population, or can it be more usefully related to males less than 5 feet tall with gray eyes and blood type A+? Such an appraisal is not often done in risk assessment and that could be improved, in as much as targeted regulations can be shown to be effective.

A whole series of other questions relates to variability modeling in PBPK models, and more precisely to model parameterization. Basically, we have two options for modeling variability. On one hand we can go for a statistical model, assuming that "people vary randomly" in such and such characteristic, and that this variability follows such and such distribution. On the other hand we may be able to reasonably assume that a physiological parameter varies with age, or physical activity, in a known (often nonlinear) manner. Yet, and that may just fall under the banner of model choice, there are several ways to model such covariate as age or physical activity. Take the example of the relationship of ventilation rate and cardiac output with activity: do we want to model it as an activity dependence of exactly those variables, or should we introduce the ventilation over perfusion ratio and keep only one of the two original variables? For blood flows to individual organs, is it better to have a Wishart-based statistical model (which is quite restrictive (Barnard *et al.* 2000)) or scale flow by organ mass and reconstruct total flow as the sum of individual organ flows? So far I have not seen any concerted, focused, attempt at dealing with those questions. Maybe that does not matter, but that is not proven either. My feeling is that introducing known covariates should be done whenever possible and should be more used than it is now.

Past and current practices

The range of practices to deal with (or simply go around) the issue of variability is very large. That may be proof that the problem has loomed large in the mind of modelers and risk assessors:

- Safety factors, as I understand them, were first linked to an interesting and very simple statistical reasoning: "We might err because human are not all the same; We are better off on the right (or left) tail of the distribution; We have 10 fingers and if we were to earn 10 times less money than we do now we would be very poor and if we earned 10 times more we would be pretty well off; therefore a factor 10 up (or down) should be enough in that case". That makes sense to almost any one. Only recently have we tried to backup those with actual data on variability and use percentiles of empirical variability distributions instead of arbitrary factors.

- Scenarios can also be defined that span the range of possibilities. Usually worst-case scenarios are used. Those are actually fine when they show no risk, but they maybe hard to define in high-dimensional nonlinear problems. The problem is that when they point to a risk the actual range of risks values is the subject of abundant controversies and cannot usually be solved without moving to more sophisticated approaches.

- Monte Carlo simulations can also be performed (Portier and Kaplan 1989). Actually scenarios as defined above can be conceived as discrete (often boundary) points in the parameter space or interest. Monte Carlo simulations simply sample many of such points according to an implicitly or explicitly defined probability distribution. They are more costly to run than sets of scenarios but they avoid the representativity problem. Often, the sampling distributions used mix uncertainty and variability, but attempts to differentiate between the two have been proposed in the form of "2D simulations" (Bogen and Spear 1987) or more sophisticated population models. An interesting form of uncoupling is found in attempts to first estimate an "average" model and then use *in vitro* data to define distributions that reflect mainly variability. However, defining an average model without a proper statistical framework is difficult, and disentangling variability from uncertainty in *in vitro* data can be quite difficult.

- Population models (Sheiner and Ludden 1992) are setup to disentangle (as much as the data allow it!) variability from uncertainty. Those are statistical models, typically hierarchical: The parameters (*e.g.*, mean) of the distribution of some variables depend on other parameters, which are themselves random variables. For example, the estimated value of a clearance parameter for a given subject will have for mean the "true value", and that value for that subject will be further distributed around a population mean, *etc*. But they are only models, and as such they embody assumptions, which can and should be criticized. For example, if we say that the body weight of subject is normally distributed in the US population we are probably wrong, and known strata exist by age, sex, ethnicity, *etc*. Those models are, however, quite convenient, because some of their parameters directly quantify variabilities and others (or their distributions) quantify uncertainty.

Current best practices

We need to be able to characterize the impact of variability on risk assessment predictions. A best practice would first offer an assessment of inter- and eventually intra-individual differences, and also be able to add estimates of the uncertainty affecting such variability and the residual uncertainty. At this moment it seems difficult to do that properly without the help of a hierarchical population model. I do not mean to imply that such a model should always be calibrated using hierarchically organized datasets (groups of individuals, observed at several occasions at which a whole set of kinetic data are collected). Bayesian (Gelman *et al.* 1996), or maximum likelihood (see M. Davidian, in this issue) methods are available for that. But one can devise a hierarchical model just for predictive purpose, using predefined (in jargon prior) distributions at all levels. Indeed these distributions should be firmly data-based and I have the feeling that human cohort data will always have a role to play, but they could also be based on *in vitro* evidence, quantitative structure-properties modeling and the like. So to be clear: the current best practice uses a hierarchical model, either in an inference-prediction cycle, or in a purely predictive mode. Even better practice should focus on setting up mixes of deterministic and stochastic modeling to capture the complex evolution of covariances between parameters in a given subject; but that maybe more on the side of research activities.

Needs

Here also potential improvements crop up as soon as we start examining the underlying hypotheses of our current practices. If we try to go from the simplest to the wildest ideas:

- We talk much of inter-individual variability. Isn't intra-individual variability important? How much does it contaminate our estimates of variability and uncertainty? Have we looked enough in that issue, and are even the data necessary to call a judgment available?
- Reparameterizing for predictions: Assume that we have obtained (through population modeling inference or *in vitro* data) a reasonable distribution for “young healthy adult males Caucasian volunteers”. We probably do not want to simply use those distributions for a realistic mixed

population, not to mention susceptible populations. Mezetti et al. (2003) have examined the inter-population difference problem for butadiene, but from the point of view of inference rather than from that of prediction. But here again I think that the solution probably calls for large datasets collected on very different individuals (but shall we ever be able to observe children and pregnant women?) and/or a deterministic modeling of inter-population differences. On that last point, recent models of the time-evolving individual are quite interesting and should be further developed and validated.

- PBPK modeling shows that inter-individual variability does exist in animals and affect internal dose. Therefore the usual statistical model in cancer dose-response analyses should *not* be binomial (or Poisson when approximated), but rather a mixture of binomials. That should apply to other dose-response models and it definitely calls for better coupling of PBPK and DR (or BBDR) models and inference. Indeed, along the same vein, what about inter-individual, inter-experiment, inter-lot, inter-lab, inter-strain variability? Do we have a sense of their relative importance? Should they be minimized to the highest possible extent? When they exist, are they amenable to meaningful statistical modeling (meta-analysis…), or should we resort to deterministic modeling as much as possible?

- Maybe, for those of us who cannot afford designing a brand new 10000 subjects study for each chemical investigated, it might be interesting to define reference distributions for inter-individual differences in major pharmacokinetic parameters. Such work has been started by Hattis *et al.* (2003), for example. As a general guideline, such reference distributions should be conditioned (stratified) on obvious covariates (age, and sex in particular). Further work on non-obvious covariates and modifiers should be encouraged.

- Extending the above allusion to modifiers, a useful tool would be afforded by a cartography of metabolic networks' (phenotypic and genotypic) polymorphism in human populations. That is obviously a distant goal, but the synergy with pharmacovigilance (*i.e.*, post-marketing drug toxicity monitoring) is obvious and is a good omen for that endeavor.

- Talking between us, statisticians, modelers and toxicologists, about population variability issues is all fine, but we should have a thought for our colleagues, epidemiologists, for whom that is also a foundational problem. The stereotypical risk assessment climax where toxicological evidence is confronted its epidemiological equivalent looks too much like one of those cruel "corrida" in which the Beast must die and Man survive! Actually the analogy runs quite deep. In any case, we should go beyond these disciplinary and cultural barriers and work on the consolidation of epidemiological and toxicological evidence. That involves collaboration with epidemiologists (who are quite interested in internal dose reconstruction), to arrive at coherent models (in particular at the dose-response level), study designs, statistical analyses, weighting of evidence *etc.*

- Have we seriously thought about the question of variability in model structure? What if the kinetics of *X* in subject *A* are mono-compartmental and for subject *B* tri-compartmental? How much evidence of that is there? Have we looked at it with adequate data? Taking a larger PBPK model and just making parametric assumptions might solve it all. However, nullity assumptions can be tricky in a probabilistic context), and we may want to use simpler models for numerical efficiency (for example).
- How far will go inter-individual heterogeneity assessment and subsequent risk estimation? How finely will regulation apply? This may be well beyond the scope of this meeting, but regulators should be aware that we might sooner or later develop, even involuntarily, the tools needed to adjust regulation (protection) down to the subject level; the same as medicine is striving to tailor therapeutic treatment to the patient level. It seems to me that some ethical and legal thinking seems in order to clarify that issue.

4. Quality assurance issues

This paper is in fact almost entirely devoted to quality assurance (QA) in PBPK model applications and corresponding guidelines development. But some more generic aspects may be worth mentioning here.

Past and current practices

Past applications of PBPK modeling in a regulatory context are again quite scarce, and most of the available works are in fact scientific publications or pre-publications, which adopt the same style. The (quite high) standard is therefore that of scientific papers, where clarity, precision and referencing are the norm. A problem stems however from the nowadays ferociously imposed brevity of the reports. Authors have to resort to supplemental material whose access can be quite problematic (not included in standard subscription, available in electronic formats whose permanence or simply longevity is far from guaranteed, *etc.*)

Current best practices

At this point in time, it seems advisable that the following be documented:

- Model structure and assumptions underlying it. Obviously referencing of previous careful work is advisable (as there is for example, little use in going over each assumption embodied in the Michaelis-Menten model of enzyme kinetics). The model equations, and if possible computer code, should be available. It might be advisable to document the proper behavior of that code in standard settings (for example, checking mass conservation), but this should be the topic of further discussion and will not be developed here.

- Choice of parameter values and/or distributions. Opting for a single value instead of a distribution (except for mathematical or physical constants) should be justified by global uncertainty analysis, with respect with the goals of the work through sensitivity analysis (see "Current best practices" in section 2, above).

- Procedures for model checking; in particular with cross-validation if model calibration has been performed with some datasets. The procedures to accept or reject the model should have been defined *a priori*. The results should be clearly and extensively documented (the creative use of graphs is recommended) (Gelman *et al.* 2002).

- Presentation and, specifically for this section on traceability and QA, storage of the predictions in ways that allow uncertainty analysis to the carried through in the context of larger models (*e.g.*, coupled PBPK and BBDR models). Obviously, the most compact way to meet this requirement in the case of statistical distributions is to check that analytical descriptions apply and report meta-parameters. But the multi-dimensional results of numerical simulations are usually not amenable to such simplifications and large sample storage is required, unless repro-modeling (which can be quite efficient) is applied.

- Documented attempts to identify the major sources of uncertainty and variability in model predictions. Formal global sensitivity analysis seems to be the best to recommend here.

- Eventual examination of the impact of alternative model structures, if they cannot be ruled out on the basis of prior knowledge. This may seems like a tall order, given the efforts that developing a single model require, but the work can be divided (as it usually is) between different modeling teams. That may put a slant toward a "competing" models approach, but an interesting "model fusion" approach has been taken by the US EPA after it funded work on trichloroethylene PBPK modeling by two different teams (Hack *et al.* 2006). At least a common ground for the development, checking, and reporting of the various models would help fuse or compare them.

Needs

- Widely accepted, flexible, state of the art, guidelines on traceability and model adequacy checking (including some standard benchmark tests (*e.g.*, mass balance, asymptotic behavior, *etc.*)

- Official, free access, software and data repositories (*e.g.*, for toxicokinetic datasets, historical data on animal physiological characteristics, *etc.*)

- A valid and accepted way to capitalize population pharmacokinetic analyses and *in vitro* data gathering (*i.e.,* the question of the use of historical data in PBPK modeling). Does it make sense to start a new analysis and model development when a new dataset is published? How to check for congruence of the new dataset with the older ones? How to decide which one to keep, or how to weight them? If compatible, or weighted, how to merge them in a new inference? In theory, the Bayesian framework is available to answer these questions (for an example, see Micallef *et al.* in press), but is it far from being fully operational.

- A way to encourage both the respect of QA guidelines and scientific creativity…

5. Presentation/communication of results

Presentation clarity is definitely linked to quality assurance in that it depends on a well-documented description of what was done and why (*i.e.*, traceability). To avoid redundancy between the above section and this one, we will rather focus here on issues related to *ease of understanding* of the results. First, let's admit it, PBPK models are complex and almost indecipherable to anyone not used to them (take for example, any of your first year PhD students). Stephen Hawkins has been reported to say that the readership of a book is halved by each equation included. That's probably true for reports too. In addition, most people do not have a clear idea of what a probability or risk is. In that context, how do we best communicate stochastic results on PBPK model predictions?

Past and current practices

If we observe with the necessary critical eye the PBPK model publications that come across for peer review, we cannot but notice the following:

- Model structure (equations) is usually well documented. However, a current trend (again linked to space shortage) is to cascade references of model descriptions, with undocumented (untracked?) changes between versions, and sometimes accumulations of errors as time goes by. The results become impossible to check and reproduce, and even the mere structure of the model can become obscure. This, obviously, does not facilitate the presentation of results.

- Illustrations of model fit are usually partial. That's human, I guess. We are proud of our models, we believe in them, and when space is short we present the good parts of the fit to a dataset, not the worst. Actually, if the worst was "good" according to some accepted criteria, that would be quite a strong statement, but the reality is that no one does that (unless I am mistaken). It seems that some general picture of the fit (log-scale all data *vs.* predictions plot) is a good thing to provide, together with some illustration (as extensive as possible) to the longitudinal time-course data and associated fits (with some indication of their range, *e.g.*, point-wise 95% confidence limits).

- Reassuring the reader/user that the model can be trusted when making predictions outside of the domain in which it has been calibrated (fitted) is a must, I think. Somehow, people have gotten it that models can go very wrong. I am not sure who put that in their head (government figures on income growth and the like, probably), but it is a lingering notion. The worst data is usually taken as more reliable that the best model! Presenting results of model checking through cross-validation is therefore definitely encouraged, if it understandable (the amount of pedagogy required should not be underestimated).
- When is comes to pure model predictions, confidence intervals should *always* be included (otherwise this meeting will be pointless). Graphs contrasting uncertainty and variability are also welcome and not difficult to produce.

Current best practices

Those can easily be interpolated from above, but the best practice would seem to be to seek feedback early on the presentation of the results. I do not want to go as far as suggesting to pass our manuscripts in our families asking our parents "do you understand what is my point in that paper?", but maybe some idea of the sort would be useful. We probably too much focus on *peer* review and should maybe pay more attention to *client* needs (in an extended sense).

Needs

May be those should be written by "clients", precisely and ideas from their community would be welcome, but as providers the following needs emerge:

- We lack widely diffused standards, or at least good examples, of "good" model description, parameters reporting, uncertainty assessment for predictions, *etc.* for selected audiences. Indeed such standards or guidelines may give a winning point in a court case, or other contradictory settings, and could be viewed as "trade secrets", but governments may want to set some for the purpose of efficient communication with stakeholders.
- A question arises as whether publication standards in toxicology are high enough in terms of thoroughness, precision, completeness in the presentation of modeling exercises.
- We might need to value "humility" when expounding domain of applicability claims for our models. I don't think that we should give the impression that we will save the world or even the day with them.

We might love to believe that, but simply contributing rationally to honest debates about what are *in fine* political choices, might be enough for as an image and presentation goal.

- One way to honestly win the sympathy and adhesion of our fellow humans could be to show them the feed back loop between experiments (or field studies) and models (theory). Obviously, decision needs to be taken, but in addition to making a decision, for "here and now", the door could be left open for a reassessment when further data is available. Modeling, *via* sensitivity analysis *etc.*, can help define which data is most useful in that context.

6. Conclusions

The real frontier may well be or, even, may have always been on model structure. Coupling PBPK models with population statistical models was an advance on structure, which now permits to better take data into account and make better predictions. A challenge is now to incorporate the wealth of *in vitro* data that *omics* can bring us. It seems to me that PBPK models are an unrecognized form of "system biology" models, even though it is one of the earliest attempts to model the body as a whole. In any case, our work on uncertainty and variability estimation will not be lost, far from that, in a system biology dominated future.

A second point is that we are usually reluctant to admit we are unsure of a model structure, and may not agree between us on it. That seems, and it may well be, more damning than being ignorant of the exact value of an obscure parameter in an otherwise clear and well-established model or theory. How can we best regulate in a presence of model uncertainty? Should we wait until the controversy is cleared, or should we act without waiting, at least in a transition phase, according to the Precautionary Principle? In that transition, is model "averaging" better than selecting the "best" model? The issue is pervasive at the interface between science and government (see economic forecasting or global climate models, for example) and health risk assessment cannot escape it.

This paper is intended to be a support for discussion and this section will probably be further enriched at the issue of group discussions. It is a sign of good health for a research field and its applications to be the topics of constructive debates.

References

Abbas, R. and J. W. Fisher (1997). A physiologically based pharmacokinetic model for trichloroethylene and its metabolites, chloral hydrate, trichloroacetate, dichloroacetate, trichloroethanol, and trichloroethanol glucuronide in B6C3F1 mice. Toxicology and Applied Pharmacology 147: 15-30.

Andersen, M. E. (1995). What do we mean by ... dose? Inhalation Toxicology 7: 909-915.

Barnard, J., R. McCulloch, et al. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. Statistica Sinica 10: 1281-1311.

Bischoff, K. B., R. L. Dedrick, et al. (1971). Methotrexate pharmacokinetics. Journal of Pharmaceutical Sciences 60: 1128-1133.

Bogen, K. T. and R. C. Spear (1987). Integrating uncertainty and interindividual variability in environmental risk assessment. Risk Analysis 7: 427-436.

Bois, F. Y. (2000). Statistical analysis of Fisher et al. PBPK model of trichloroethylene kinetics. Environmental Health Perspectives 108 (suppl. 2): 275-282.

Bois, F. Y., M. Smith, et al. (1991). Mechanisms of benzene carcinogenesis: application of a physiological model of benzene pharmacokinetics and metabolism. Toxicology Letters 56: 283-298.

Bois, F. Y., T. Smith, et al. (1999). Optimal design for a study of butadiene toxicokinetics in humans. Toxicological Sciences 49: 213-224.

Clewell, H. J., G. P.R., et al. (2000). Development of a physiologically based pharmacokinetic model of trichloroethylene and its metabolites for use in risk assessment. Environmental Health Perspectives 108 (suppl. 2): 283-305.

Davidian, M. and A. R. Gallant (1992). Smooth nonparametric maximum likelihood estimation for population pharmacokinetics, with application to quinidine. Journal of Pharmacokinetics and Biopharmaceutics 20: 529.

Droz, P. O., M. M. Wu, et al. (1989). Variability in biological monitoring of organic solvent exposure - II. Application of a population physiological model. British Journal of Industrial Medicine 46: 547-558.

Droz, P. O., M. M. Wu, et al. (1989). Variability in biological monitoring of solvent exposure - 1. Development and validation of a population model. British Journal of Industrial Medicine 46: 447-460.

Fisher, J. W. (2000). Physiologically based pharmacokinetic models for trichloroethylene and its oxidative metabolites. Environmental Health Perspectives 108 (suppl. 2): 265-273.

Frederick, C. B., M. L. Bush, et al. (1998). Application of a hybrid computational fluid dynamics and physiologically based inhalation model for interspecies dosimetry extrapolation of acidic vapors in the upper airways. Toxicology and Applied Pharmacology 152: 211-231.

Gearhart, J. M., D. A. Mahle, et al. (1993). Variability of physiologically based pharmacokinetic (PBPK) model parameters and their effects on PBPK model predictions in a risk assessment for perchloroethylene (PCE). Toxicology Letters 68: 131-144.

Gelman, A. (2004). Parameterization and Bayesian modeling. Journal of the American Statistical Association 99: 537-545.

Gelman, A., F. Y. Bois, et al. (1996). Physiological pharmacokinetic analysis using population modeling and informative prior distributions. Journal of the American Statistical Association 91: 1400-1412.

Gelman, A., X.-L. Meng, et al. (1996). Posterior predictive assessment of model fitness via realized discrepancies. Statistica Sinica 6: 733-807.

Gelman, A., C. Pasarica, et al. (2002). Let's practice what we preach: Turning tables into graphs. American Statistician 56: 121-130.

Gentry, P. R., T. R. Covington, et al. (2003). Evaluation of the potential impact of pharmacokinetic differences on tissue dosimetry in offspring during pregnancy and lactation. Regulatory Toxicology and Pharmacology 38: 1-16.

Georgopoulos, P. G., A. Roy, et al. (1994). Reconstruction of short-term multi-route exposure to volatile organic compounds using physiologically based pharmacokinetic models. Journal of Exposure Analysis and Environmental Epidemiology 4: 309-328.

Hack, C. E., W. A. Chiu, et al. (2006). Bayesian population analysis of a harmonized physiologically based pharmacokinetic model of trichloroethylene and its metabolites. Regulatory Toxicology and Pharmacology 46: 63-83.

Hattis, D., G. Ginsberg, et al. (2003). Differences in pharmacokinetics between children and Adults - II. Children's variability in drug elimination half-lives and in some parameters needed for physiologically-based pharmacokinetic modeling. Risk Analysis 23: 117-142.

Jonsson, F. and G. Johanson (2001). Bayesian estimation of variability in adipose tissue blood flow in man by physiologically based pharmacokinetic modeling of inhalation exposure to toluene. Toxicology 157: 177-193.

Leahy, D. E. (2006). Integrating in vitro ADMET data through generic physiologically based pharmacokinetic models. Expert Opinion on Drug Metabolism & Toxicology 2: 619-628.

Mezzetti, M., J. G. Ibrahim, et al. (2003). A Bayesian compartmental model for the evaluation of 1,3-butadiene metabolism. Journal of the Royal Statistical Society Series C-Applied Statistics 52: 291-305.

Micallef, S., B. Amzal, et al. (in press). Sequential updating of a new dynamic pharmacokinetic model for caffeine in premature neonates. Clinical Pharmacokinetics.

Müller, P. (1999). Simulation-based optimal design. Bayesian Statistics 6. J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. mith. Oxford, Oxford University Press: 459-474.

Occupational Safety and Health Administration (OSHA) (1997). Occupational exposure to methylene chloride - Final rule. Federal Register January 10 1997, 29 CFR Parts 1910, 1915, and 1926: 1494-1611.

Poet, T. S., R. A. Corley, et al. (2000). Assessment of the percutaneous absorption of trichloroethylene in rats and humans using MS/MS real-time breath analysis and physiologically based pharmacokinetic modeling. Toxicological Sciences 56: 61-72.

Portier, C. J. and N. L. Kaplan (1989). Variability of safe dose estimates when using complicated models of the carcinogenic process. Fundamental and Applied Toxicology 13: 533-544.

Reitz, R. H., J. F. Quast, et al. (1980). Non-linear pharmacokinetic parameters need to be considered in high dose / low dose extrapolation. Archives of Toxicology suppl. 3: 79-94.

Rosner, G. L. and P. Mueller (1997). Bayesian population pharmacokinetic and pharmacodynamic analyses using mixture models. Journal of Pharmacokinetics and Biopharmaceutics 25: 209-234.

Saltelli, A., S. Tarantola, et al. (2000). Sensitivity analysis as an ingredient of modeling. Statistical Science 15: 377-395.

Sheiner, L. B. and T. M. Ludden (1992). Population pharmacokinetics/dynamics. Annual Review of Pharmacology and Toxicology 32: 185-209.

Spear, R. C., F. Y. Bois, et al. (1991). Modeling benzene pharmacokinetics across three sets of animal data: parametric sensitivity and risk implications. Risk Analysis 11: 641-654.

Srinivasan, R. S., D. W. Bourne, et al. (1994). Application of physiologically based pharmacokinetic models for assessing drug disposition in space. Journal of Clinical Pharmacology 34: 692-698.

Zhang, P. (1993). Model selection via multifold cross validation. Annals of Statistics 21: 299-313.